ТРАНСФЕР ТЕХНОЛОГИЈЕ ЗА ОБРАДУ ПРИРОДНИХ ЈЕЗИКА ЕКСПЕРИМЕНТИ, МОГУЋНОСТИ И ОГРАНИЧЕЊА СТУДИЈА СЛУЧАЈА: ТРАНСФЕР С ЕНГЛЕСКОГ НА СРПСКИ

Мирослав Мартиновић, Колеџ Њу Џерзи, Одсек за рачунарство
Превод с енглеског: Бојана Ђорђевић

Апстракт: У овом раду описујемо истраживање и наставне активности везане за један научни експеримент. Студија је спроведена како би се испитале могућности и ограничења трансфера технологија везаних за природне језике с технолошки богатог а морфолошки скромног језика као што је енглески на језик чија је технологија у развитку а морфологија богата, као што је српски. У кратким цртама представљамо нека колаборативна достигнућа која укључују изградњу система за проналажење информација (SPretS), тагера врста речи (SPaS), плитког парсера (SPaSer), и система за праћење тематика (SeTo TraS) за српски језик. Неки од овде приказаних резултата већ су познати захваљујући радовима са конференција, док су други у процесу објављивања.

(I) Општи преглед

Рад који овде описујемо представља део Сабатикал пројекта, замишљеног као низ истраживачких активности удружених са предавањима на лицу места или он-лајн предавањима на постдипломском нивоу или нивоу основних студија у циљу обрађивања одређених тема.

Истраживачке активности се групишу око ресурса и алата из области рачунарске лингвистике, пре свега оних које сам лично развио за енглески језик (Програм за разрешавање анафора 'AARLISS' ([35]), програм за препознавање именованих ентитета 'HyNER', програм за лематизацију/конфлацију 'SteLemMin' ([37]), упитнички систем 'QASTIIR' ([34])). Изучена је објављена литература и испробани су, анализирани и процењени сродни алати. Локални истраживачи су касније подстакнути да обаве сродно истраживање на тему развоја ресурса за рачунарску лингвистику за српски језик.

Предавања су обухватала преглед стања савремене методологије и достигнућа у области обраде природних језика, с нагласком на она која се односе на мале језике попут српског. Теорија и пракса претраживања и проналажења текстуалних и библиографских информација представљени су с посебним нагласком на проблеме везане за одговарање на питања. Осим тога, темељно су проучени савремени и најуспешнији приступи проблему идентификовања међу бројним он-лајн текстуалним документима одговора на питања постављена на природном језику, како би се поставио темељ за текућа и даља истраживања.

(II) Претходна истраживања

Поље технологије људских језика коначно је довољно сазрело те се много већа пажња посвећује развоју лингвистичких ресурса за језике којима говори мање од десет милиона људи (мале језике). Развој многих (усмерених) апликација попут командовања и управљања, диктирања, препознавања говорне поште, аудио претраге, аутоматског превођења новинских текстова и скупштинских записника достигле су висок квалитет погодан за практичну употребу. Системи природних језика проширени су на већи круг језика. ЕУ сада има 23 званична језика и користи већи број машински читљивих речника (МЧР) (за нпр. фински, чешки, мађарски, словеначки).

Истраживачки пројекти широких размера у САД такође усмеравају пажњу на МЧР за мале језике (нпр. DARPA GALE ([18]), Transtac ([19]): за модерни стандардни арапски, арапске дијалекте, фарси, DARPA CAST: за пашто). Започет је низ нових пројеката у циљу развоја језичких ресурса (нпр. NEMLAR ([20]): мрежа за евромедитеранске језичке ресурсе; TELRI ([21]): трансевропска инфраструктура језичких ресурса усмерена на језике централне и источне Европе; LDC пројекат језика који се ређе уче (LCTL) ([17]): развија језичке ресурсе за урду, таи, мађарски, бенгали, панџаби, тамилски и јорубу). Уведени су нови нивелациони задаци и процене како би се мерио напредак (*Morpho* изазов: за процену алгоритама за аутоматску морфолошку сегментацију и анализу; CoNLL дељени задатак: концентрише се на депенденцијално парсирање за већи број језика (укључујући чешки, арапски, турски, баскијски, мађарски).

Осим тога, и етнолингвистичка анализа односа између културе, мисли и језика, а посебно изучавање мањинских језика у контексту већинске популације, већ неко време добија значајан подстрек ([3], [4], [10], [11], [12], [13], [14], [15],[16], [22], [26], [27], [29], [36], [42]). Започете су и промовишу се стандардизација, лингвистичка нормализација и ревитализација малих језика, о чему сведочи и повећан број веб-страна на мање коришћеним језицима.

Аутоматска обрада малих језика треба да превазиђе низ потекоћа које проистичу из њиховог особеног статуса.

(а) Како ови језици имају мало матерњих говорника, имају и мало лингвиста којима је то матерњи језик и још мање рачунарских лингвиста. Због тога може доћи до потешкоћа у примењивању традиционалних приступа тагирању, парсирању итд. који се заснивају на правилима.

(б) Слаба финансијска подршка коју ови језици добијају изгледа да такође практично искључује приступе базиране на правилима, с обзиром на количину људског рада коју ови приступи обично захтевају. Овај проблем би се могао превазићи усвајањем рачунарских оквира изведених на основу других језика.

(в) Корпусни приступи су примењиви једино ако су доступни одговарајући корпуси. С друге стране, формирање корпуса је захтевно и у временском и у финансијском смислу, и захтева добру лингвистичку концепцију, посебно ако је у питању корпус опште намене.

(г) Приступи базирани на примерима изгледа у овом смислу више обећавају уколико су потребни специфични примери, а не корпуси опште намене. Такође се чини и да је лакше да се имплементира компилација специјалних примера него да се пишу формална правила. Ипак, мало се зна о изводљивости овог обрасца у случају мањинских језика.

(д) Могу се развити, или су већ у употреби технике плитког знања које користе специфичне особине језика или језичких породица. Ово, с друге стране, може ометати трансфер таквог приступа од једног језика ка другим језицима. Неке технике би тако могле да функционишу са аналитичким, али не и са аглутинативним језицима итд. Различити системи писања такође могу да спрече примењивост једног једноставног приступа на друге језике.

(III) Циљеви

Читавом пољу рачунарске лингвистике тренутно недостају значајнија истраживања на малим језицима. Неколицина талентованих студената и истраживача из Србије упозната је с најсавременијим истраживањима на том пољу преко курсева „Проналажење информација" и „Системи за одговарање на питања из отвореног домена". Уз изучавање препоручене литературе и демонстрацију ресурса и алата развијених кроз одговарајућа истраживања за енглески језик, ово је подстакло сарадњу на низу битних питања попут:

(а) Однос између обраде природних језика и подршке малим језицима уопште.

(б) Развој наменских апликација за обраду природних језика за српски, као што су оне за тагирање, морфолошку анализу, парсирање, проналазак информација или одговарање на питања.

(в) Развој корпуса и машински читљивих речника за српски језик.

(г) Презентација техника плитког знања за обраду природних језика које би биле примењиве на српски језик.

(д) Прегледне студије које описују стање обраде природних језика за српски и друге мале језике читавог региона или сличног језичког типа.

(ђ) Компаративна студија различитих приступа обради природних језика за различите мале језике и језичке типове.

(е) Слободни ресурси за обраду природних језика, поља њихове примене и њихова ограничења.

(ж) Захтеви апликација за обраду природних језика за српски и сродне језике.

(IV) План пројекта

У току оба семестра комбинован је низ истраживачких активности које ће бити описане са подучавањем на лицу места и он-лајн предавањима на постдипломском нивоу или на нивоу основних студија у циљу обрађивања одређене теме.

У току првог семестра предавачке активности су подразумевале курс из Система за проналажење информација, сличан оном који сам већ држао раније у више наврата. Курс је укључивао и преглед савремених методологија и достигнућа из области обраде природних језика, с нагласком на оне које се односе на мале језике попут српског. Детаљно се проучавала теорија и пракса претраживања и проналажења текстуалних и библиографских информација, с нагласком на проблеме везане за одговарање на питања. Овај курс је био спој предавања, презентација чланака, оцењивачког и практичног рада на стварању и развоју пројекта. Студенти су проучавали препоручену литературу, презентовали и критички оцењивали радове, а такође су и развијали сопствени систем за проналажење информација, што је био њихов семестарски пројекат и основа за даље истраживање. Тежило се што већој флексибилности архитектуре овог система, што је требало да омогући његову каснију трансформацију у Систем за екстраховање информација и, најзад, у Систем за одговарање на питања. У току курса, студенти су се упознали и са расположивим лингвистичким алатима за тагирање, парсирање, разрешавање референци и препознавње именованих ентитета, од којих сам неке лично развио.

На тај начин је постављен основ за следећи семестар и наредни курс на тему Системи за одговарање на питања отвореног домена. Детаљно су проучени савремени и најуспешнији приступи проблему идентификовања међу бројним он-лајн текстуалним документима одговора на питања постављена на природном језику, како би се могло прећи на даља истраживања. Као и претходни, и овај курс је био спој предавања, презентација чланака и критичког осврта на њих и практичног рада на стварању и развоју пројекта. Студенти су проучавали препоручену литературу, презентовали и критички оцењивали радове. Од студената се очекивало и да развију модуле који би касније били интегрисани у њихов сопствени систем за одговарање на питања, што је био њихов пројекат за тај семестар и основ за будуће истраживање. Студенти су у току курса упознати и са неким од постојећих система за одговарање на питања, као и са QASTIIR-ом, који сам лично развио.

Као сарадднички пројекти, потекли из пројеката овог курса, осмишљени су, промовисани и развијени пројекти који се баве специфичностима ресурса и алата за српски језик, како је и описано у даљем тексту.

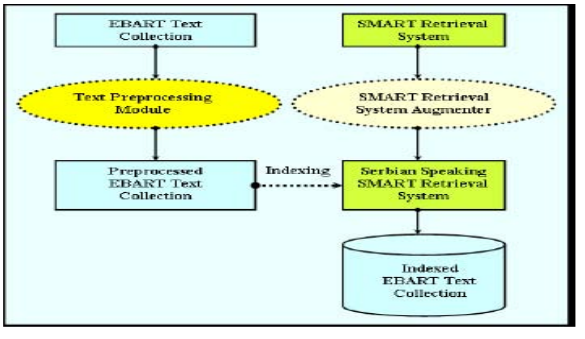
(V) Ширење и исходи

Значајнији резултати ове мисије пренети су научној јавности у оквиру четири колаборативна рада, саопштена или пријављена на разним конференцијама рачунарске лингвистике (INTERSPEECH 2007, CICLING 2008, LREC 2008).

(1) „Изградња система за проналажење информација за српски – изазови и решења", Мирослав Мартиновић, Срђан Весић, Горан Ракић (пријављен, прихваћен и представљен на конференцији INTERSPEECH 2007 одржаној у Антверпену у августу 2007).

Овај рад је описивао изазове на које се наилазило при изградњи система за проналажење информација за српски језик. Као окосница нашег система коришћен је систем за проналажење информација SMART, унапређен својствима неопходним за рад са специфичностима српског алфабета. Осим тога, морфолошко богатство језика истакло је значај припремне фазе у обради текста. У току ове фазе, конструисана су два алгоритма која су увећала прецизност проналажења за 14%, односно 27%. Тестирање је било спроведено на ЕБАРТ колекцији српских новинских чланака, величине два гигабајта.

Општа структура српског система за проналажење информација (СПретС (Српски ПРЕТраживачки Систем)) приказана је на Дијаграму 1.



Дијаграм 1: Општа структура СПРЕТС-а

И колекција текстова и систем SMART се модификују и претходно обрађују како би се процес индексирања могао правилно извести. Затим се проширени систем SMART за проналажење информација, прилагођен за обраду српског језика, примењује на претходно обрађену колекцију текстова. Индексирање које следи ствара индекс текстуалне колекције ЕБАРТ. Претходна обрада текстуалне колекције састоји се од три главне фазе: (1) конверзија не-ASCII српских слова (*ћ, ч, ђ, ш* и *ж*) у одговарајуће ASCII транскрипте (*cx, cy, dx, sx*, и *zx*), (2) одбацивање и (3) конфлација речи.

Док су конверзија не-ASCII слова и одбацивање прилично једноставне операције, за осмишљавање и развој успешног алгоритма за конфлацију се то не би могло рећи. Притом овај процес има и далекосежне последице на функционисање система за проналажење информација.

Српски језик има доста компликовану морфологију: седам именичких падежа, три именичка рода, три именичка броја, шест глаголских времена, шест глаголских форми итд, итд. ([23], [46]). Било је сасвим логично претпоставити да би квалитет конфлације могао имати битне последице на процес проналажења информација. Ово је, на крају, и доказано нашим каснијим истраживањем, као и утврђивањем и проценом учинка система.

Из тог разлога смо се одлучили, не само да развијемо систем за проналажење информација, већ и да експериментишемо са њим и да меримо утицај његових различитих конфлационих модула. У том циљу, поставили смо експеримент који се састојао из три корака у коме се покретао систем и мерио његов учинак најпре без модула за конфлацију, затим уз употребу алгоритма за подсецање с исцрпним бројем правила, и, најзад, уз употребу алгоритма за подсецање са само најосновнијим начелима. Према томе, развили смо два алгоритма за конфлацију моделира-

на према неким од добро познатих претходника на том пољу ([1],[3], [9], [36], [39], [40], [41], [51]).

Исцрпни алгоритам за конфлацију (ИАК) састоји се од педесетак трансформационих правила која укључују подсецање завршетака речи, као и анализу префикса и замену слова у средини речи. Типично правило састоји се од свог обележја, услова дужине речи (веома кратке речи се не трансформишу) и описа облика речи праћеног трансформацијом (формат је установљен за STELEMMIN, генерички минимални стемер/лематизатор ([36])). Како се трансформација може извршити било где у речи, потребна је позиција трансформације, као и опис обрасца који треба да буде замењен као и оног који долази уместо њега. Сви описи су регуларни изрази. Следи пример типичног трансформационог правила:

31; M>3; реч на претпоследњој позицији има а коме претходи консонант, а иза кога не следе **ч, ћ, ђ, ш, ж, љ, њ** *или* **ц***; одстрани а; замени га са ‘’.*

Овај метод представља покушај препознавања сличности међу речима не само када се оне разликују у роду, броју или времену, већ и када постоје сложене гласовне алтернације. У српском се јављају бројне појаве тог типа попут палатализације, преласка *л* у *о*, непостојаног *а* итд. ([23], [46]).

Правило палатализације подразумева конверзију гласова *к, г, х у ч, ж, ш,* на позицији испред вокала, (нпр. номинатив *друг* се у вокативу трансформише у *друже*).

Правило које доводи до конверзије *л* у *о* примењује се када се род придева и именица мења из женског у мушки или средњи (нпр. номинатив женског рода це*ла* и номинатив мушког рода цео).

Правило непостојаног а брише претпоследње а у речи ако му претходи консонант, а иза њега не следе **ч, ћ, ђ, ш, ж, љ, њ** или **ц** (нпр. генитив једнине *мањка* изведен од номинатива једнине именице *мањак*).

У другом примеру, реч *математички* трансформише се у *математик* најпре примењујући правило које одстрањује финално *ки*, а затим оно које замењује *ч* са *к*.

Реч *најјачи* трансформисана је у *јк* примењивањем следећих правила овим редоследом: брисање префикса *нај,* брисање вокала на крају речи, замењивање *ч* са *к*, брисање *а* на претпоследњој позицији на којој је окружено консонантима.

Због своје исцрпне природе, овај алгоритам успева да трансформише већину речи у њима одговарајуће заједничке основе. Ипак, и сасвим очекивано, запажени су и случајеви претеране нормализације. На пример, реч *пруге* (у смислу *железничке пруге)* и *пругасти* трансформишу се у исту реч *пруга* (*железничка пруга*). Због сличних случајева претеране нормализације, одлучили смо се да експериментишемо с алтернативним алгоритмом. Изостављање правила одговорних за наведене случајеве претеране нормализације довело нас је до следећег алгоритма.

Алгоритам за рудиментарну конфлацију (АРК) развијен је редуковањем претходног на само мали подскуп првобитних правила. Овде су правила која се односе на раније поменуте сложене линвистичке појаве изостављена, као и нека друга за која је утврђено да су доводила до претераног подсецања. При обради упита *пруге Србије*, ИАК је сводио речи *пруге* и *пругасте* на исти корен што је доводило до забавних али озбиљних грешака у проналажењу: пронађен је чланак који се бави популарношћу Барби лутки у *пругастим* купаћим костимима. У другом, сличном случају, за упит *кирија*, алгоритам за ИК свео је речи *кирија*, *Кира* (властито име) и *Кири* (презиме) на исту основу.

Нови алгоритам за РК не доводи до претеране нормализације и може да створи различи-

те подсечене облике у случајевима аналогним претходним (*пруге* и *пругасте*; *кирија, Кира* и *Кири*). Уопште је показао већу прецизност, што ће детаљно бити и објашњено.

У односу на претрагу без икакве претходне конфлације (наша полазна процедура), општи закључак је да су оба алгоритма довела до значајног побољшања и у одзиву и у прецизности. Док је алгоритам за исцрпну конфлацију (АИК) довео до значајног повећања одзива и квалитетног пораста прецизности, алгоритам за рудиментарну конфлацију (АРК) довео је до пристојног раста одзива и супериорног повећања прецизности.

Како је раније поменуто, избор агенције ЕБАРТ српских новинских чланака објављених 2003-2007, величине два гигабајта, коришћен је као наша колекција докумената. Учињен је напор да се испрате опште препоруке TREC упутстава за процену учинка система за проналажење информација („http://www-nlpir.nist.gov/projects/tv2007/tv2007.html"). Постављен је експеримент за процену који је мерио неинтерполирану P-прецизност у подскуповима ЕБАРТ докумената растућим по величини где је сваки следећи скуп надскуп претходног и садржи нових пет до тада не-обрађених докумената (првих 5 докумената, првих 10 докумената, првих 15, првих 20 итд). Тестирање је обављено на 106 упита скупљених од насумично изабраних корисника. Прецизност је била забележена за сваки појединачни упит на сваком од одабраних 'модуо 5' подскупова докумената. Као последњи корак, рачуната је просечна појединачна прецизност свих подскупова докумената чији су бројеви конгруентни по модулу 5. Табела 1 показује средњу вредност прецизности по скуповима докумената као и укупну средњу вредност по свим подскуповима докумената, као и процентуална увећања ових вредности за алгоритаме за ИК и РК у односу на претрагу без икакве конфлације.

Document Count	Ground Zero Precision	ECA Average Precision	RCA Average Precision
5	0.698113	0.745283	0.822642
10	0.600943	0.676415	0.754717
15	0.520161	0.632704	0.701258
20	0.478338	0.575977	0.643904
...	...	...	...
Query Average	**0.57439**	**0.65759**	**0.73063**
% Increase	-	**14.5 %**	**27.2 %**

Табела 1. P-прецизност статистика

(2) „Тагирање врста речи и плитко парсирање у језицима с мањим бројем ресурса: случај српског", Мирослав Мартиновић, Младен Николић (пријављен је и тренутно у разматрању за конференцију LREC 2008).

Овај рад представља експеримент на изграђивању тагера врста речи и плитког парсера за језик са скромним ресурсима за језичку технологију као што је српски. Замишљен је и усвојен генерички и флексибилан приступ вођен спољашњом конфигурацијом у којој се као формализам за опис фраза користе регуларни изрази. Као његова окосница коришћен је тагер TnT. С обзиром на то да у корпусу намењеном обучавању није било информација о роду и броју, за изградњу плитког парсера коришћена су само обележја врсте речи. Као корпус је коришћена анотирана ЕБАРТ колекција српских новинских чланака. Тестирање наших алата показало је да је њихов учинак био упоредив са одговарајућим алатима за језике за које постоје богати ресурси.

Тагирање врстама речи (граматичко тагирање или на енглеском *POS tagging* или POST) је процес анотирања речи у тексту на основу тога којој врсти речи прпадају. Тагови

(етикете) се бирају и на основу дефиниције речи и на основу њеног контекста (односа са суседним и повезаним речима у фрази, реченици или пасусу). Тагирање врстом речи се данас скоро искључиво обавља у контексту рачунарске лингвистике. Алгоритми који су у употреби повезују засебне термине, као и скривене врсте речи, у складу са скупом описних тагова.

Плитко парсирање (познато и као подсецање или „лако парсирање") је анализа текста у току које се идентификују његови конституенти (именичке фразе, глаголске фразе, предлошке фразе, итд). Међутим, од овог процеса се не очекује ни да спецификује унутрашњу структуру нити улогу тих фраза у главној реченици. Ово је процедура која се у великој мери користи у обради природних језика.

Осмишљавање и развој тагера SPaS (од енглеског *Serbian Part of Speech*) започети су прибављањем висококвалитетног српског корпуса анотираног врстама речи. Ово је покренуло низ даљих радњи, вођених нашим избором да користимо језички независан тагер TnT. Први корак је била проста коверзија корпуса ЕБАРТ у формат који прихвата генератор модела TnTGenerator. Потом је генератор параметара тагера обучаван на претходно обрађеном и анотираном корпусу. Резултат је било стварање модела параметара формирањем два нова ресурса: лексичких и контекстуалних фреквенција за српски корпус. Ово је уједно био и крај фазе компилације. Два новостворена ресурса и компонента за тагирање TnT-а чине стварну оксоницу система СПаС за тагирање. У време извршавања, нови корпус (који тек треба да буде тагиран врстама речи) треба најпре прерадити у формат који захтева TnT систем. Он се затим, заједно са лексичким и контекстуалним фреквенцијама створеним у току фазе компилације, уноси у компоненту за тагирање TnT-а. Коначан резултат је текст тагиран врстама речи.

Нацрт SPaSer-а (од енглеског *Shallow Parser for Serbian*) је развијен као генерички и високо флексибилан алат за обраду језика. Покреће га конфигурацијска поставка која се уноси споља, а која садржи дефиниције за све фразе које могу бити парсиране. Фразе се дефинишу коришћењем регуларних израза преко врста речи. Дефиниције фраза могу да садрже и раније уведене фразе (које се означавају унутар дефиниције ограђивањем симболом '#').

Садржај ове конфигурацијске датотеке, која се користи за дефинисање најпростијих именичких фраза (if – од именска фраза), предлошких фраза (pf – предлошка фраза) и глаголских фраза (gf – глаголска фраза) у SPaSer-у, изгледа овако:

```
<vrste_reci>
<vrsta id="1">imenica</vrsta>
<vrsta id="2">pridev< /vrsta>
<vrsta id="3">broj</vrsta>
<vrsta id="4">zamenica</vrsta>
<vrsta id="5">glagol</vrsta>
<vrsta id="6">prilog</vrsta>
<vrsta id="7">predlog</vrsta>
<vrsta id="8">veznik</vrsta>
<vrsta id="9">uzvik</vrsta>
<vrsta id="0">recca</vrsta>
<vrsta id=".">tacka</vrsta>
<vrsta id=",">zarez</vrsta>
<vrsta id="!">znak_uzvika</vrsta>
<vrsta id="?">znak_pitanja</vrsta>
<vrsta id="(">otvorena_zagrada</vrsta>
<vrsta id=")">zatvorena_zagrada</vrsta>
<vrsta id="nov_red">nov_red</vrsta>
<vrsta id="default">ostalo</vrsta>
</vrste_reci>
<fraze>
<fraza id="if" komentar="Imenske fraze">
        (prilog*pridev+)*imenica+
    </fraza>
    <fraza id="pf" komentar="Predloske fraze">
        predlog #if#
    </fraza>
    <fraza id="gf" komentar="Glagolske fraze">
        glagol #if#
    </fraza>
</fraze>
```

SPaSer очекује улазну колекцију, као и XML конфигурацијску датотеку. Како је раније поменуто, у конфигурацијској датотеци се преко регуларних израза дефинишу фразе које могу бити парсиране. Улазни текст се затим тагира врстама речи помоћу компоненте тагера SPaS који у реалном времену заправо обавља анотацију. Када се у улазном тексту нађу све информације о врстама речи, конфигурацијски изрази се интерпретирају како би се препознало присуство фраза у тексту које се могу парсирати. Коначно, излаз се састоји од плитко парсираног текста обележеног у XML-у.

Фрагмент текста парсираног SPaSer-ом изгледа овако:

```
<if>
<imenica>godini</imenica>
</if>
<prilog>znatno</prilog>
<gf>
<glagol>popraviti</glagol>
<zamenica>njihov</zamenica>
<if>
<imenica>standard</imenica>
</if>
</gf>
<zarez>,</zarez>
<gf>
<glagol>pokazuje</glagol>
<if>
<imenica>istraživanje</imenica>
<imenica>agencije</imenica>
<imenica>Pressing</imenica>
</if>
</gf>
```

Процену рада тагера SPaS мерили смо на основу његове тачности (однос исправно тагираних речи према свим тагираним речима). Анотирани корпус је, на самом почетку, био подељен на део за обуку и онај за тестирање (након што су уклоњени сви тагови). Тачност SPaS-ових тагова мерена је у односу на одговарајуће ручно постављене тагове у првобитном скупу за тестирање. Процена је урађена за различите пропорције података за обуку и сумирана је у Табели 1.

Training Data %	0.001 %	0.01%	0.1%	1%	10%	50%	90%
Test Data Accuracy	53.6%	71.8%	85.8%	93.1%	97.1%	98.5%	98.9%
% of Known Words	26.9%	40.7%	59.5%	79.8%	92.6%	96.7%	97.7%
Accuracy on Known Words	99.6%	99.7%	99.4%	99.4%	99.5%	99.5%	99.5%
Accuracy on Unkn. Words	36.7%	52.6%	65.7%	68.2%	66.6%	69.8%	71.9%

Табела 1. Табела процене за тагер SPaS

Табела укључује и тачност за скуп за тестирање, као и проценат речи у корпусу за тестирање које су већ биле познате из корпуса за обуку, укључујући прецизност и за познате и за непознате речи. Укупан број речи у читавом корпусу био је око 11.000.000.

Учинак плитког парсера процењиван је мерењем нивоа прецизности и одзива. Прецизност је процењивана коришћенем два приступа. У првом, који се најбоље може описати као ‘бинарни’, добија се оцена 1 за сваку идентификовану максималну фразу, а затим се рачуна укупни проценат у односу на читав скуп фраза. Алтернативни метод оцењивања, који се може описати као ‘неизразити’ (од енглеског *fuzzy*), даје оцену за било који број речи из фразе који систем идентификује. Ако се, на пример, идентификује фраза која се састоји од две речи, али максимална фраза, чији је ова само један део, садржи пет речи, даје се оцена 0,4 (или 2/5). Оба приступа кажњавају присуство речи које не припадају фрази давањем оцене 0,0.

Одзив се одређује на аналогни начин, с том разликом што се не узимају у обзир све фразе које је систем идентификовао, већ оне које је требало да идентификује. Текст коришћен за тестирање састојао се од 142 именске фразе.

	Binary Scoring	Fuzzy Scoring
Precision	0.79	0.87
Recall	0.82	0.88

Табела 2. Процена одзива/прецизности за плитки парсер

Наш плитки парсер је очигледно заиста легитиман алат. У случајевима када је циљ налажење фразе, а не њене максималне варијанте, SPaSer се може са сигурношћу окарактерисати као веома успешан.

Приметили смо још једну интересантну чињеницу: није било скоро никакве разлике у учинку када је парсер обрађивао фразе просечне дужине (пет или шест речи) и веома дуге фразе (девет или десет речи).

(*3) „Векторизација структуираног текста са неизразитом евалуацијом за праћење теме у језицима с мањим бројем ресурса: случај српског", Мирослав Мартиновић, Душан Васић (пријављен је и тренутно се разматра за конференцију LREC 2008).*

Овде се бавимо проблемом моделирања праћења теме у језицима попут српског. Усвојени приступ проистиче из класичног векторског модела. Запазили смо да новински чланци на вебу по правилу имају сложену структуру (датум, аутор, наслов, често и поднаслов и наднаслов могу се пронаћи, а често су и опште теме као што су политика, спорт и култура директно доступни). Подесили смо систем за проналажење SMART тако да искористимо претпостављену минималну структуру текста помоћу технике структуиране векторизације. У току процедуре учења, добијене су оптималне тежине за различите делове чланка. Тестирање и процена су се базирали на чињеници да се косинусна функција векторског модела може користити за директно дефинисање неизразитог (fuzzy) односа толеранције. Цена класичне теорије праћења била је затим апстрахована на неизразити ниво и било је дозвољено његово специјализовање за цене односа. Корпус коришћен за тестирање и процену састојао се од чланака српских дневних новина „Политика", доступних на вебу.

Одлучили смо да систем и алгоритам, који су представљени у овом раду, назовемо SeToTraS (од енглеског *Serbian Topic Tracking System*). Систем је у потпуности имплементиран у Јави и садржи пакете развијене за објектно-оријентисану подршку решавању општих проблема векторског моделирања праћења теме. До развоја и имплементације овог система дошло је након студије и процене примењивости векторски моделираних система на откривање и праћење теме у контексту српских текстова на вебу. Приступ заснован на моделу векторског простора ([2], [5], [44]) прихваћен је као полазна тачка због своје једноставности, ефикасности и језичке независности, као и подесности коју је показао при изградњи SPretS-a, система за проналажење информација за српски језик.

Идентификовање недостатака тренутно водећег векторског модела у овим оквирима, подстакло је развој новог алгоритма који је довео до побољшања која обећавају.

SeToTraS је први, нови и једини у потпуности изграђен и комплетан систем за праћење теме за српски.

Испитивали смо концепт праћења теме усвајањем модела векторског простора као окоснице нашег система, као и неизразиту (fuzzy) логику за процену.

Успели смо да раздвојимо сакупљање корпуса, њихову организацију, обраду термина, означавање текста, векторску нормализацију, анотационе табеле, тестирање и процену алгоритма и машинско учење. На тај начин су све важније одлуке биле раздвојене, како би се олакшале било какве будуће модификације, оптимизације, или чак револуционарни преокрети у односу на елементе пројекта. Тако, на пример, наш алгоритам за сакупљање корпуса скида одређене веб стране, екстрахујући притом структуру чланка и податке, али се он

може заменити било којим другим аналогним алгоритмом (нпр. оним базираним на SQL-у, у случају базе чланака).

Применили смо само најосновније алгоритме за подсецање и векторску нормализацију. Оба захтевају даља побољшавања, као што је једноставно замењивање подсецања лематизацијом. Општа искуства са системима за проналажење информација сличним SMART-у могла би такође бити од помоћи при оптимизацији овог последњег елемента.

(VI) Захвалност

Као део Сабатикал истраживања на тему могућности и ограничења при трансферу технологија за обраду природних језика са ресурсима богатих на ресурсима мање богате језике, рад на овом пројекту делимично је финансиран једнократном стипендијом фондације Студеница, као и једнократном стипендијом Светске универзитетске службе.

Хтели бисмо да се захвалимо и колеџу Њу Џерзи на обезбеђивању логистичких и финансијских могућности за ово истраживање, као и Математичком факултету Универзитета у Београду који је великодушно угостио рад на овом пројекту. Хтео бих да изразим своју посебну захвалност Др Витасу и Др Крстев на њиховој спремној помоћи и саветима.

Литература

1. Abney, S.: Part-of-Speech Tagging and Partial Parsing. In: Church, K., Young, S., Bloothooft, G. (eds): Corpus-Based Methods in Language and Speech (1996)

2. Allan, J.: Modeling Topics for Detection and Tracking. In: Pattern Recognition in Speech and Language Processing, Chou, W., Juang, F. (eds.), CRC Press (2002) 349-372

3. Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D., Harper, D. J., Hiemstra, D., Hofmann, T., Hovy, E., Kraaij, W., Lafferty, J., Lavrenko, V., Lewis, D., Liddy, L., Manmatha, R., McCallum, A., Ponte, J., Prager, J., Radev, D., Resnik, P., Robertson, S., Rosenfeld, R., Roukos, S., Sanderson, M., Schwartz, R., Singhal, R., Smeaton, A., Turtle, H., Voorhees, E., Weischedel, E., Xu, J., Zhai., C. Challenges in Information Retrieval and Language Modeling. SIGIR Forum, March 2003.

4. Alemu, A., Asker, L., Getachew, M. Natural Language Processing for Amharic: Overview and Suggestions for a Way Forward, in Proceedings of TALN 2003 Workshop on Natural Language Processing of Minority Languages and Small Languages, June, 2003.

5. Ariki, Y., Takao, S.: Study on New Term Weighting Method and New Vector Space Model Based on Word Space in Spoken Document Retrieval. In: Proceedings of the International Conference on Recherche d'Informations Assistee par Ordinateur (RIAO'00) 4 (2000) 116-131

6. Beigbeder, M., Mercier, A.: An Information Retrieval Model Using the Fuzzy Proximity Degree of Term Occurrences. In: Proceedings of the 2005 ACM symposium on Applied computing (2005) 1018 – 1022

7. Brants, T.: TnT -- a Statistical Part-of-Speech Tagger. In: Proceedings of the 6th Applied NLP Conference, ANLP-2000, (2000)

8. Cameron, R.D.: REX: XML Shallow Parsing with Regular Expressions. In: CMPT TR 1998-17, School of Computing Science, Simon Fraser University (1998)

9. Chrupala, G.: Simple Data-Driven Context Sensitive Lemmatization. In: Proceedings of SEPLN 2006 (2006)

10. Cucerzan, S., Yarowsky, D.: Bootstrapping a Multilingual Part-of-Speech Tagger in One Person-Day. In: Proceeding of the 6th Conference on Natural Language Learning Vol. 20 (2002) 1-7

11. Echihabi, A., Oard, D.W., Marcu, D., Hermjakob, U.: Cross-Language Question Answering at the USC Information Sciences Institute. In Proceedings of CLEF 2003: Cross-Language Evaluation Forum. Workshop No4 Vol. 3237 (2003) 514-522

12. Fagundes da Silva, C., Osório, F.S., Vieira, R. Evaluating the Use of Linguistic Information in the Pre-processing Phase of Text Mining, in Proceedings of TALN 2003 Workshop on NL Processing of Minority Languages and Small Languages, June, 2003.

13. Fellbaum, C., editor, WordNet, An Electronic Lexical Database. MIT Press, 1998.

14. Gasperin, C., Vieira, R., Goulart, R., Quaresma, P. Extracting XML Syntactic Chunks from Portuguese

Corpora, in Proceedings of TALN 2003 Workshop on NL Processing of Minority Languages and Small Languages, June, 2003.
15. Hajic, J., Cmejrek, M., Dorr, B., Ding, Y., Eisner, J., Gildea, D., Koo, T., Parton, K., Penn, G., Radev, D., Rambow, O. Natural Language Generation in the Context of Machine Translation. Technical Report, Center for Language and Speech Processing, JHU, 2002.
16. Hauptmann, A., Scheytt, P., Wactlar, H., Kennedy, P.E.: Multi-Lingual Informedia: A Demonstration of Speech Recognition and Information Retrieval across Multiple Languages. In: Proceedings of the DARPA Workshop on Broadcast News Understanding Systems (1998)
17. http://projects.ldc.upenn.edu/LCTL/
18. http://www.darpa.mil/ipto/Programs/gale/index.htm
19. http://www.darpa.mil/ipto/Programs/transtac/index.htm
20. http://www.nemlar.org/
21. http://www.telri.ac.uk/
22. Hughes, B.: Towards a Web Search Service for Minority Language Communities. In: Proceedings of Open Road Conference (2006)
23. Ивић, П., Пешикан, М., Клајн, И., Брборић, Б.: Српски језички приручник, Београдска књига, Београд (2007)
24. Jesperson, O., Language, its Nature, Origin and Development, George Allen & Unwin, London, 1921.
25. Kinyon, A.: A Language-Independent Shallow Parser Compiler. In: Proceedings of 10th EACL Conference (2001) 322-329
26. Korenius, T., Laurikkala, J., Jarvelin, K. and Juhola, M. "Stemming and Lemmatization in the Clustering of Finnish Text Documents", Proceedings of the 13th ACM International Conference on Information and Knowledge Management, Session IR-7, pp. 625-633, 2004.
27. Kraaij, W. and Pohlmann, R., "Porter's Stemming Algorithm for Dutch", Noordman LGM and de Vroomen WAM, eds. Informatiewetenschap 1994: Wetenschappelijke bijdragen aan de derde STINFON Conferentie, Tilburg, pp. 167-180, 1994.
28. Kraaij, W. and Pohlmann, R. "Evaluation of a Dutch Stemming Algorithm" Rowley J, ed. The New Review of Document and Text Management, Vol. 1, Taylor Graham, London, pp. 25-43, 1995.
29. Lam, W., Chan, K., Radev, D., Saggion, H., Teufel, S. Context-based Generic Cross-lingual Retrieval of Documents and Automated Summaries. Journal of the American Society for Information Science and Technology 56(2), February 2005.
30. Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., Thomas, S.: Relevance Models for Topic Detection and Tracking. In: Proceedings of the Human Language Technology Conference (HLT) (2002) 104-110
31. Leroy, G., Chen, H., Martinez, J. D.: A Shallow Parser Based on Closed-Class Words to Capture Relations in Biomedical Text. In: Journal of Biomedical Informatics 36 (2003) 145-158. 12
32. Leuski, A., Allan, J.: Improving Realism of Topic Tracking Evaluation. In: Proceedings of ACM Conference on Research and Development in Information Retrieval (2002) 89-96
33. Li, X., Roth, R.: Exploring Evidence for Shallow Parsing. In: Proceedings of the Annual Conference on Computational Natural Language Learning (2001)
34. Martinovic, M., "Integrating Statistical and Linguistic Approaches in Building Intelligent Question-Answering Systems", Proceedings of the SSGRR 2002 International Conference on Advances in Infrastructure for e-Business, e-Education, e-Science, and e-Medicine on the Internet, 2002.
35. Martinovic, M., Curley, A., Gaskins, J. AARLISS – an Algorithm for Anaphora Resolution in Long-distance Inter Sentential Scenarios, In Proceedings of the 8th International Conference on Speech, Text and Dialogue, September 2005.
36. Martinovic, M., and Rofrano, L. "SteLemMin – A Generic Minimal Stem Algorithm for Word Conflation and Lemmatization", Proceedings of Workshop on Computational Modeling of Lexical Acquisition, 2006.
37. Martinovic, M., Sampath, G., Wagner, R., Briening, S. A Multilevel Text Processing Model of Newsgroup Dynamics : Implementation and Results, In Proceedings of the 8th International Conference on Applications of Natural Language to Information Systems, NLDB'2003, 168-175, June 2003.
38. Martinovic, M., Vesic, S., Rakic, G.: Building an Information Retrieval System for Serbian – Challenges and Solutions. In: Proceedings of the 8th Annual International Interspeech Conference (2007)
39. Miller, G. WordNet : A Lexical Database for English. In C ACM, 38(1):49-51, 1995.
40. Paice, C. D., "Another Stemmer", ACM SIGIR Forum, Vol. 24, Issue 3, pp. 56-61, 1990.
41. Porter, M. F., "An Algorithm for Suffix Stripping", Program Vol. 4, No. 3, pp. 130-137, 1980.
42. Radev, D.R., Brew, C. editors. Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Philadelphia, PA, 2002.

43. Ratnaparkhi, A.: A Maximum Entropy Model for Part-Of-Speech Tagging. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (1996)
44. Salton, G. (ed): The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall (1971)
45. Schmechel, N.: On the Lattice-Isomorphism between Fuzzy Equivalence Relations and Fuzzy Partitions. In: Proceedings of ISMVL-95 (1995)
46. Стевановић, М., Савремени српскохрватски језик, I, II, Београд, 1994.
47. Thiele, H.: On the Mutual Definability of Fuzzy Tolerance Relations and Fuzzy Tolerance Coverings. In: Proceedings of the 25th International Symposium on Multiple-Valued Logic (1995) 140
48. Thiele, H., Schmechel, N.: On the Mutual Definability of Fuzzy Equivalence Relations and Fuzzy Partitions. In: Proceedings of the International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium (1995)
49. Voorhees, E. M., "Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness", Information Processing and Management, 36(5), pp. 697-716.
50. Wayne, C. L.: Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation. In Proceedings of LREC2000 (2000).
51. Xu, J., Croft, W.B. (1998) Corpus-based Stemming using Co-occurrence of Word Variants. In ACM Transactions on Information Systems 16(1), pp. 61-81. 1998.